Accounting for deterministic noise components in a MMSE STSA speech enhancement framework

Matthew McCallum and Bernard Guillemin
Dept. of Electrical and Computer Engineering, The University of Auckland
Auckland, New Zealand
Email: {m.mccallum,bj.guillemin}@auckland.ac.nz

***Abstract*—Current approaches to speech enhancement usually consider performance in either the presence of coloured broadband noise or periodic noise, but rarely both. In this paper we present a new speech enhancement technique, derived within the standard minimum mean-square error (MMSE) short time spectral amplitude framework, which jointly compensates for both coloured broadband and periodic noise components. With this approach, termed noise mean subtracted (NMS) MMSE, hidden Markov model based frequency tracking techniques are used to estimate periodic noise components and differentiate them from periodic components associated with the speech signal. They are then removed using complex spectral subtraction. The resulting algorithm is evaluated using perceptual evaluation of speech quality (PESQ) for both male and female speech utterances from the NOIZEUS database. It is shown that when the noise contamination comprises both broadband and periodic components, this NMS MMSE algorithm outperforms the standard MMSE algorithm, derived under the assumption of stochastic noise only. The technique has application in a variety of scenarios, including those involving emergency radio communications.**

***Index Terms*—Minimum mean-square error (MMSE), speech enhancement, deterministic noise, emergency communications.**

I. Introduction

Various mobile speech communications devices are now in widespread use. The ability to communicate accurately and reliably on such devices is of great interest to a range of users. Enabling reliable communication for emergency services personnel is a particularly difficult problem due to two major counteracting factors. First, communication for these personnel is potentially very urgent, therefore, information must be communicated in a precise and robust manner. Secondly, the environments in which these personnel communicate can have significantly adverse effects on a speech signal's quality and intelligibility due to prominent background acoustic noise.

Of the several major communication problems emergency service personnel have with regards to acoustic noise, one of particular interest in this paper is the presence of fire truck pump noise in fire fighter communications. This interest is due to the consistency with which this problem occurs in a fire fighter's work, the particularly low signal to noise ratio (SNR) that is often encountered, and finally the generally unexploited structure of fire truck pump noise.

This research is supported by Capability funding from the MBIE Science and Innovation Group.

A common communication scenario fire fighters encounter involves one member of the fire fighter team controlling the operation of the fire truck water pump. Whilst doing so, this member must constantly communicate with other members on the scene who are often out of their line of sight. Communicating in close proximity to the fire truck water pump is made difficult by the loud acoustic noise it generates.

The signal to be enhanced in such a scenario is a single-channel noise corrupted speech signal. A range of fundamental speech enhancement approaches have been proposed in the literature [1], [2], [3], [4], all of which aim to exploit specific characteristics of the speech signal such as its quasi-stationarity or its harmonic structure. A broad review on the performance of many of these fundamental approaches can be seen in [5], [6]. While many refinements have been made upon these approaches that allow for modelling of the speech signal with more desirable or accurate characteristics [7], [8], and consideration of the human auditory perceptual system, [9], [10], there has been much less attention paid to how these speech enhancement algorithms may be tailored for improved performance in the presence of specific noises. While noise from some noise sources fit very well into the original assumptions made in the development of certain speech enhancement algorithms, many noise sources produce noise that has characteristics that do not agree with these assumptions. The time-frequency structure seen in fire truck pump noise is an example.

The time-frequency structure of fire truck pump noise appears to be a combination of broadband and sinusoidal components, and is seldom addressed directly in speech enhancement. This is surprising due to its common occurance in radio communication scenarios. Recently such a structure was addressed with respect to adaptive filtering speech enhancement techniques which have been known to be effective in mitigating periodic noises [11]. These speech enhancement techniques are less widely researched than the range of broadband speech enhancement techniques covered in [12]. Due to their thorough research in the literature, the theory upon which broadband speech enhancement techniques are developed is well understood and furthermore, many problems specific to speech enhancement have already been addressed for such techniques. While some broadband speech enhancement techniques will improve audio quality in the presence of periodic noises [1], [2], [3], it is important to consider what further

gains are possible when periodic noise structure is considered explicitly. This may allow improved reduction of periodic noise in speech signals, whilst maintaining the major advances already made in these speech enhancement techniques. In this paper the time-frequency structure of fire truck pump noise is addressed under the framework of broadband speech enhancement techniques. Furthermore, similar structures are observed in a range of other noises, indicating this is an important general problem in speech enhancement. Through this investigation a new speech enhancement algorithm is developed, termed the noise mean subtracted minimum mean square error (NMS MMSE) algorithm, that offers improved speech enhancement performance in the presence of such noises.

This paper is organised as follows, Section II describes the time-frequency structure of fire truck pump noise and highlights the particular features of this structure that are not currently exploited in most single channel speech enhancement algorithms. Section III describes how this time-frequency structure may be appropriately exploited in a broadband speech enhancement algorithm. Estimation of the instantaneous parameters specifying the structure of fire truck pump noise is discussed in Section IV. Finally, in Section V, a speech enhancement algorithm is investigated with and without a component that works to exploit this structure, revealing the potential of such an algorithm in the presence of this particular noise.

II. Description of the Considered Noise Type

The structure of fire truck pump noise is of particular interest with regards to speech enhancement because it does not adhere to the assumptions of noise statistics that are made in the development of most current speech enhancement algorithms. There are a range of well known speech enhancement methods in the literature that pertain to white and coloured broadband noise [1], [2], [13], and periodic noise [14]. However, none of these noise characterisations appear to apply when analysing the power spectrogram of fire truck pump noise. Figures 1(a), 1(b) and 1(c) provide three examples of fire truck pump noise. Upon observing the spectrograms here, it appears that this noise cannot be accurately classed as purely broadband or purely periodic noise, but it is an additive mixture of both. For example, paying specific attention to Fig. 1(a) there is clearly broadband energy across all frequencies here, but at least four higher energy sinusoidal components are observed, starting at approximately 200Hz, 1.25kHz, 2kHz and 2.5kHz, and increasing in frequency thereafter. Considering the fire truck pump mechanism that is creating this acoustic noise, this idea seems reasonable. The water pump mechanism consists of several rotating components which are likely to create high energy periodic acoustic signals. In addition, the flow of fluids both within the pump, and around such rotating components, is likely to produce turbulence, resulting in noise of a more stochastic nature.

Analysis of such mixed signals is not a new concept and thorough reviews can be found in [15], [16]. In such a case, the signal is considered to consist of several deterministic sinusoids composing the periodic components and a stochastic component that may have energy at all frequencies, including those of the sinusoids. This stochastic/deterministic noise signal may be modelled as follows:

$$d[n] = \sum_{l=1}^{L} r_l \cos\left(f_l n + \phi_l\right) + h[n], \tag{1}$$

where L is the number of sinusoidal components in the noise model, and for each component l, r_l is the amplitude, f_l is the normalized frequency in radians per sample and ϕ_l is the

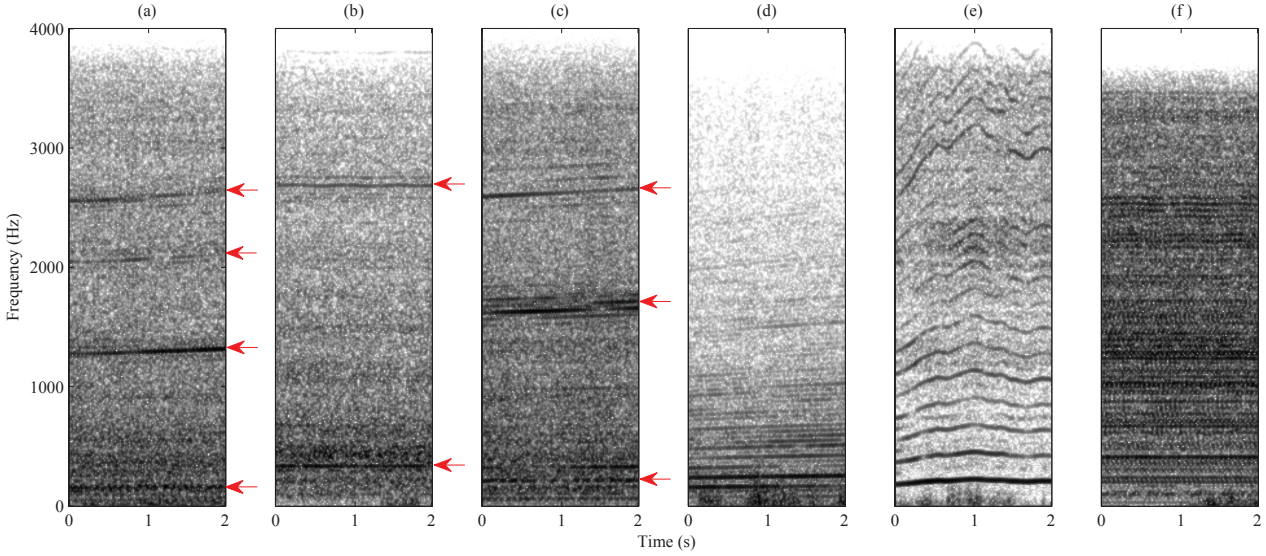

Figure 1. Spectrograms of a range of noises. (a)-(c) Fire truck pump noise samples, (d) boat (predominantly outboard motor) noise, (e) chainsaw noise, (f) fire truck engine noise. In figures (a) through (c) regions of spectrum containing dominant deterministic noise components have been labelled explicitly with arrows. Such components in figures (d) through (f) are also clear, but too numerous to warrant explicit labelling. All spectrograms were obtained with a Hamming window of length 512 samples, and an overlap of 75%. All noise samples were sampled at a rate of 8kHz.

phase. Finally, $h[n]$ is a stochastic process characterised by its power spectral density (PSD). Equation (1) will be referred to as the stochastic/deterministic (SD) noise model throughout the remainder of this paper.

As will be discussed in Section III, the SD noise model allows a tractable modification to certain speech enhancement frameworks that accommodates for these observed properties of the noise explicitly. Although the SD noise model has been defined here with respect to fire truck pump noise, a wide range of noise sources consist of rotating and turbulent mechanisms that produce noise similar to that produced by fire truck pumps. Therefore, this noise structure is also applicable in a more general sense. For example, a wide range of motors such as those in boats and chainsaws, those in various items of factory machinery and those in certain household appliances, all involve rotating and turbulent mechanisms that are likely to result in similar noise structures. This idea is supported upon observing the power spectrograms emitted by a few of these noise sources as seen in Figures 1(d), 1(e) and 1(f).

III. Deterministic noise components in speech enhancement

As mentioned in Section I, a speech enhancement method involving the use of adaptive filters has been proposed to work effectively in the presence of noises representable by the SD noise model [11]. However, it would be imprudent to ignore the development of the wider class of speech enhancement algorithms [12] for such common noises. In particular, minimum mean square error short-time spectral amplitude (MMSE STSA) speech enhancement algorithms [1] are widely known to provide promising performance [6], [5]. Furthermore, this method is based on the well researched principles of spectral analysis [16]. Hence, in these algorithms, the inclusion of the SD noise model described in Section II is not only tractable, but allows the use of a set of thoroughly developed spectral analysis tools. Due to the capable performance of such a speech enhancement algorithm and its relevance to the particular problem which this paper addresses, the MMSE STSA algorithm in [1] is investigated here with respect to the SD noise model.

Many speech enhancement algorithms consider the enhancement of a noisy speech signal in terms of a short-time framework. This is motivated by the quasi-stationarity assumed in both the speech and noise signals. Under this framework, the digital signal

$$y[n] = x[n] + d[n] \tag{2}$$

is analysed over the interval $[0, N]$ for which stationarity can be assumed, where $x[n]$ and $d[n]$ represent the speech and noise processes, respectively. Typically N is a number of samples corresponding to a timespan in the order of 10-30ms (i.e., 80-240 samples at a sampling rate of 8kHz). This signal is then windowed with an appropriate windowing function, $w[n]$, and transformed via the discrete Fourier transform (DFT) to yield a set of Fourier coefficients,

$$Y_k = B_k e^{j(k+\beta_k)} = \frac{1}{T+1}\sum_{n=0}^{N} w[n]y[n]e^{-j2\pi\frac{k}{N}n}. \tag{3}$$

Similar representations may be considered for speech and noise spectra, $X_k = A_k e^{j(k+\alpha_k)}$ and $D_k = C_k e^{j(k+\psi_k)}$, respectively. Here k refers to the discrete frequency bin index. The variables B_k, A_k and C_k refer to the noisy speech, clean speech and noise amplitudes, respectively, and β_k, α_k and ψ_k refer to the noisy speech, clean speech and noise phases, respectively. In brief, the derivation of a MMSE STSA speech enhancement algorithm involves considering how the coefficients, X_k, D_k and hence Y_k, are distributed and applying Bayes' theorem to find the MMSE estimate of the clean speech spectral amplitude A_k, given the observations Y_k, under the assumed distributions. In the presence of the noises shown in Fig. 1, the distribution typically assumed for the noise in MMSE STSA speech enhancement algorithms [1] is not entirely applicable. Under these conditions the SD noise model in (1) is more appropriate as it explicitly accounts for the periodic components indicated in Fig. 1. In the presence of this noise model, the noise signal in the STFT domain may be represented for a given frame as,

$$D_k = \sum_{l=1}^{L} \frac{r_l}{2}\left(e^{j\phi_l}W_{f_l,k} + e^{-j\phi_l}W_{-f_l,k}\right) + H_k \tag{4}$$

$$= \sum_{l=1}^{L} Q_{l,k} + H_k \tag{5}$$

where $W_{f_l,k}$ and H_k, for all k, are the set of Fourier coefficients resulting from the STFT of the frequency shifted windowing function, $w[n]e^{jf_l n}$, and the stochastic process, $h[n]$, respectively. Whilst H_k is considered a zero-mean, complex Gaussian distributed random variable [1], the variables $Q_{l,k}$ may be considered to be deterministic parameters. Due to this deterministic component, the noise Fourier coefficients, D_k, may be considered to be complex Gaussian distributed with a non-zero mean [15], and hence their probability density function (pdf) may be expressed as,

$$p(D_k; \mu_k, \rho_k) = \frac{1}{\pi\lambda_{D_k}} \exp\left\{-\frac{1}{\lambda_{D_k}}|D_k - \mu_k e^{j\rho_k}|^2\right\}, \tag{6}$$

where the parameters μ_k and ρ_k define the amplitude and phase of the non-zero mean, respectively. This distribution is shown graphically in Fig. 2. Here the distribution is seen to be a two dimensional function of the real and imaginary parts of the Fourier coefficients, D_k. The dependence on the complex mean, $\mu_k e^{j\rho_k}$, is explicitly labelled. This is a deterministic parameter and it may be considered to be the contribution to the Fourier coefficient at index k due to sinusoidal components in the noise signal spectrum.

To incorporate this statistical characterisation into the MMSE STSA algorithm, U_k may be defined as,

$$U_k = Y_k - \mu_k e^{j\rho_k}. \tag{7}$$

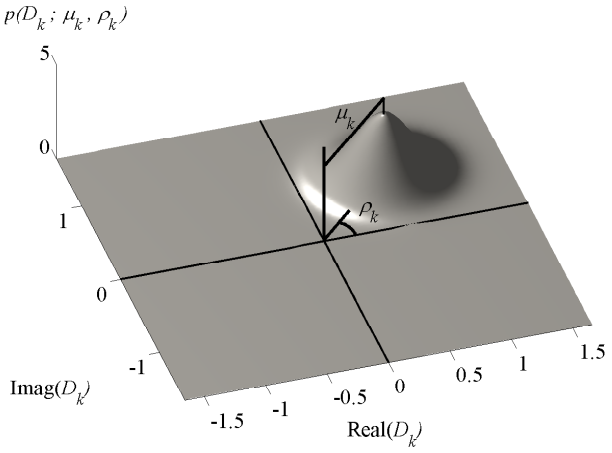

Figure 2. The distribution of complex DFT noise samples under the assumed SD model. The mean magnitude, μ_k, and phase, ρ_k, are labelled explicitly.

Then considering $Y_k = X_k + D_k$, the conditional probability,

$$p(X_k|U_k) = \frac{1}{\pi\lambda_{D_k}} \exp\left\{-\frac{1}{\lambda_{D_k}}|U_k - X_k|^2\right\}. \qquad (8)$$

may be derived. This expression is seen to be identical to that assumed in the MMSE STSA speech enhancement algorithm described in [1]. However, the direct spectrum observation Y_k is now modified according to (7), and so the observation, U_k, in (8), is better thought of as the noise-mean subtracted (NMS) spectrum.

In practice, μ_k and ρ_k are unknown and must be estimated from the noisy speech signal. An appropriate estimation procedure is briefly described in Section IV. Once estimated, the parameters, L, r_l, f_l and ϕ_l can be substituted into (1). By taking the DFT of this equation for $0 \leq n \leq N$, the noise mean spectrum $\mu_k e^{j\rho_k}$ for frequency bin k may be obtained. Next, the NMS spectrum is calculated via (7). The MMSE STSA estimator in the presence of a SD noise model is then achieved by applying the MMSE STSA gain function ((7) of [1]), to the NMS spectrum of the noisy speech signal. The resultant complete system is appropriately named NMS MMSE and can be seen in Fig. 3. The modification made to the MMSE STSA speech enhancement system, accounting for the SD noise model, is labelled "Sinusoidal noise component compensation". This involves the deterministic parameter estimation procedure described in Section IV.

IV. Deterministic parameter estimation

From (7) it is clear that there are several parameters that must be estimated to compute the NMS spectrum. These include the set of frequencies, amplitudes and phases that comprise the deterministic part of the spectrum, $\mu_k e^{j\rho_k}$, and also the discrete PSD characterising the stochastic process, $h[n]$.

Estimating the frequency of sinusoidal components in a signal given a finite observation window is a well known problem for which many possible solutions exist [16]. Because the

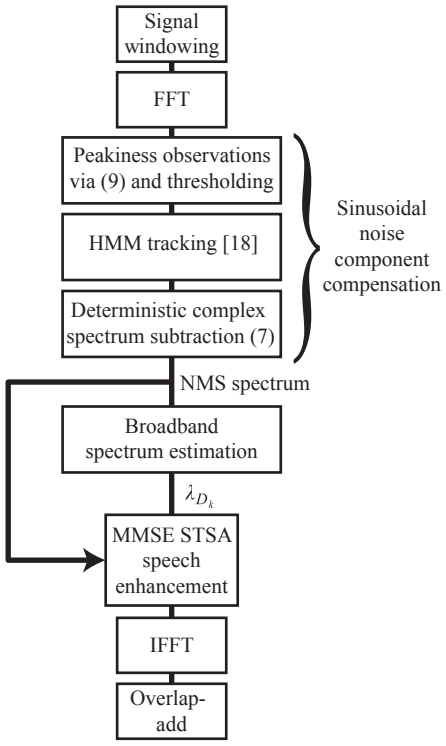

Figure 3. A flow diagram detailing the major processes in the NMS MMSE speech enhancement system detailed in this paper. Here the signal flow is from top to bottom.

sinusoidal components in fire truck pump noise are not necessarily harmonically related, more sophisticated techniques that exploit harmonic relationships between sinusoidal components of a signal are not relevant. Through experimenting with a range of frequency estimation methods, we have found that the maximum-likelihood method [16] is a relatively reliable frequency estimation technique for fire truck pump noise. For sinusoidal signals with L sinusoidal components, where the relationship between sinusoidal components is not harmonic and is unknown, the maximum-likelihood method of frequency estimation corresponds to picking the L maxima from the magnitude discrete time Fourier transform of the signal. Given that in this case L is also unknown it is proposed here that instead, local spectrum maxima are selected based on the crest factor criteria introduced in [17],

$$P_k[\kappa] = \frac{||C_{sub}[k]||_\infty}{||C_{sub}[k]||_2}. \qquad (9)$$

$P_k[\kappa]$ may be thought of as the ratio of peak spectral magnitude to RMS spectral magnitude and is indicitive of the local "peakiness" of the signal magnitude spectrum. The vector, $C_{sub}[k]$, is a subinterval of the noise magnitude spectrum, C_k, centered at frequency bin κ. That is, the elements in $C_{sub}[k]$, correspond to the indices

$$\kappa - (\Omega - 1)/2 \leq k \leq \kappa + (\Omega - 1)/2 \qquad (10)$$

of C_k, where Ω is the peakiness criteria bandwidth in terms

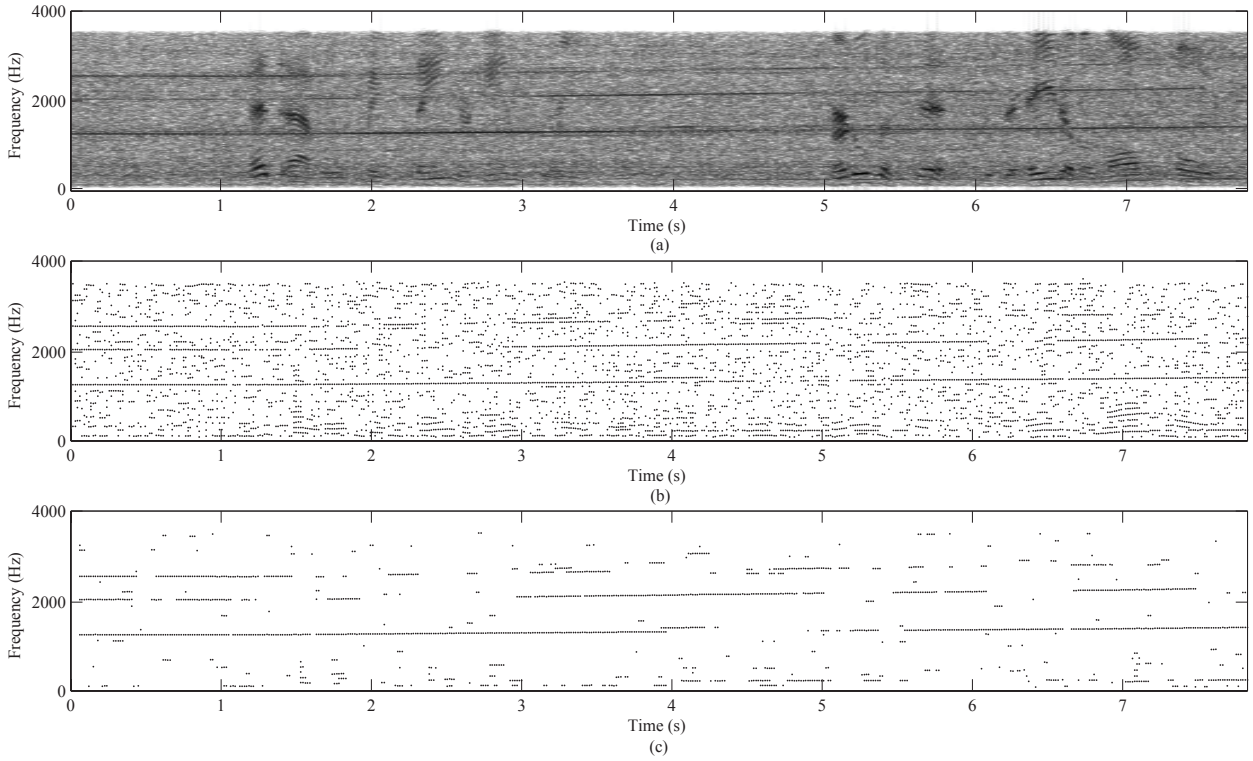

Figure 4. (a) A spectrogram of speech additively mixed with fire truck pump noise at an SNR of 0dB. (b) A set of pitch observations obtained via the thresholded crest factor measure of (9) for $P_k[\kappa] > 2$ and $\Omega = 28$. (c) The resultant HMM tracks of major sinusoidal components that have a likelihood greater than 0.2. The disruption of the pitch tracks in this figure is representative of the fading in and out of determinstic components in the fire truck pump noise signal. In all figures here, the signal is sampled at 8kHz, a windowing length of 60ms is used with 75% overlap and an FFT length of 3840 samples.

of a number of DFT bins. $P_k[\kappa]$ may be evaluated at all maxima of C_k, although a more selective choice of maxima may improve computational complexity.

In identifying sinusoidal components of the speech signal displayed in Fig. 4(a) (corrupted with fire truck pump noise). Estimates may be obtained by thresholding solutions to (9) accross all spectral magnitude peaks. The set of estimates obtained is seen in Fig. 4(b). It is clear here that there are many spurious peaks. Furthermore, in order to compute the NMS spectrum, spectral peaks that exist due to the periodicity in a speech signal must be differentiated from those that compose the unwanted noise signal. Upon comparing a speech signal and a fire truck pump noise signal, a major differentiating factor between the periodicities in each is that frequencies of sinusoids in fire truck pump noise change less often and at a slower rate than those in a speech signal. With the use of a hidden Markov model (HMM), it is possible to impose the expected time-varying behaviour of a sinusoid's frequency upon a set of observed frequency estimates, to obtain a most likely sequence of estimates, as described in [18]. Using this technique and paying specific attention to the likelihood of each of the resultant frequency estimates it is possible to identify sequences of estimates that best match the time-varying properties defined in the HMM's specification. For example, in the case of fire truck pump noise, the HMM transition pdf may be set to closely represent frequencies that change only by small amounts or, more often, not at all between succesive STFT windows. Therefore sequences based on only more slowly varying observations will have a higher likelihood, and may be identified as components of the fire truck pump noise.

The methods described here are shown in context of the NMS MMSE algorithm in Fig. 3, where they comprise the blocks in the "Sinusoidal noise component compensation" section. The output of this section of the algorithm is seen in Fig. 4(c), where it is clear that they are able to detect most of the obvious deterministic noise components accurately in frequency and time, and reject the vast majority of spurious spectral peaks seen in Fig. 4(b). The remaining noise signal parameters $\mu_k e^{j\rho_k} = \sum_{l=1}^{L} Q_{l,k}$ and H_k can be estimated via standard spectral estimation techniques [15], [16].

V. Experimental results

The NMS MMSE system seen in Fig. 3 was tested over a variety of speech utterances combined with fire truck pump noise at a range of SNRs. These results were compared with the standard MMSE STSA speech enhancement system, identical to that described in [1], using the perceptual evaluation of speech quality (PESQ) metric [19]. The results of these tests are seen in Fig. 5.

A VAD based broadband spectrum estimation component was used in both systems [20]. Windowing lengths of 60ms and 30ms, and overlaps of 50% and 75% were used for the deterministic parameter estimation described in Section IV,

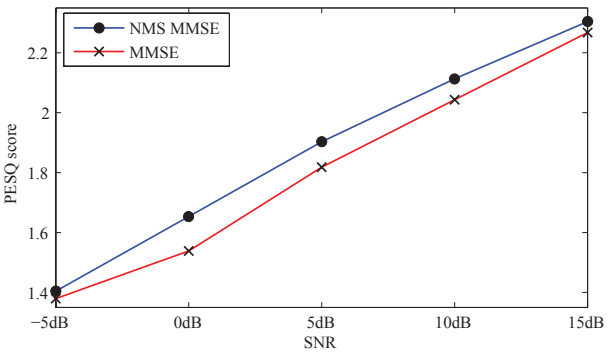

Figure 5. PESQ averages of speech enhancement results with the original MMSE STSA algorithm [1] (labelled MMSE) and the algorithm specified in this paper (labelled NMS MMSE).

and the STFT processing in the MMSE STSA algorithm [1], respectively. The longer windowing length generally obtains more accurate results for the estimation of deterministic noise components due to their slowly varying nature.

The MMSE STSA speech enhancement algorithm requires a parameter most commonly referred to as α, that controls the amount of dependence on *a priori* information from previous frames [1]. For all experiments here $\alpha = 0.98$. Speech utterances were taken from the NOIZEUS database [5]. A total of 20 utterances, 10 male and 10 female, from four different speakers, were combined and prepared adhering to the recommendations in [19]. Each was combined with a randomly selected sample of recorded fire engine pump noise at specific SNRs (-5dB, 0dB, 5dB, 10dB and 15dB), using the methods in [21]. The PESQ scores of tests at each SNR were then averaged to produce a final result. The results of these tests show that the NMS MMSE algorithm outperforms the standard MMSE algorithm at all SNRs.

VI. Conclusion

This paper has presented a new MMSE STSA speech enhancement system, named NMS MMSE, that is derived based on a noise model that allows for both deterministic and stochastic noise components. These deterministic components are seen in a variety of common noise sources as shown in Fig. 1, with the specific case of fire truck pump noise investigated in this paper. The explicit estimation of deterministic noise components has been successfuly achieved using the techniques described in [15], [16] and [18]. In particular, the slowly varying nature of deterministic components in fire truck pump noise is exploited. The extension of this speech enhancement algorithm to the other noises demonstrated in Figures 1(d), 1(e) and 1(f) will require consideration of how deterministic components are distributed in frequency and time. Specifically, aspects such as the harmonic relationship of deterministic components, the spectral envelope of deterministic components and the typical patterns resulting from changes in the frequency of these components may be useful considerations. Once estimated, the deterministic components are compensated for using complex spectral subtraction.

Such a system results in more effective attenuation of these deterministic components than the original MMSE STSA speech enhancement system, which is derived based only on a stochastic noise model. The system described here is important for the attenuation of noises in, for example, critical emergency services communications. In particular, it is experimentally proven to be effective in the presence of fire truck pump noise.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[3] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[4] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 4, pp. 251–266, 1995.

[5] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7, pp. 588–601, 2007.

[6] ——, "A comparative intelligibility study of single-microphone noise reduction algorithms," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, 2007.

[7] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 845–856, 2005.

[8] Y. Hu and P. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 1, pp. 59–67, 2004.

[9] P. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 857–869, 2005.

[10] P. Wolfe and S. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *Acoustics, Speech, and Signal Processing. IEEE International Conference on*, vol. 2, 2000, pp. 821–824.

[11] K. Sumi, N. Sasaoka, Y. Itoh, K. Fujii, and S. Tsuiki, "Noise reduction method for sinusoidal and wideband noise based on ALE and noise reconstruction filter," in *Communications and Information Technology. IEEE International Symposium on*, vol. 1, 2004, pp. 567–570.

[12] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[13] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 4, pp. 334–341, 2003.

[14] R. Sherratt, D. Townsend, and C. Guy, "Cancellation of siren noise from two way voice communications inside emergency vehicles," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 4, 1999, pp. 2395–2398.

[15] D. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 1055–1096, 1982.

[16] S. M. Kay, *Modern spectral estimation: theory & application*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[17] S. Boyd, "Multitone signals with low crest factor," *Circuits and Systems, IEEE Transactions on*, vol. 33, no. 10, pp. 1018–1022, 1986.

[18] M. Wu, D. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 3, pp. 229–241, 2003.

[19] ITU-T, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, 2001.

[20] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[21] ITU-T, *Objective measurement of active speech level*, ITU-T Recommendation P.56, 2011.